Customizable Asymmetric Loss Functions for Machine Learning-based Predictive Maintenance

Lukas Ehrig
Hasso Plattner Institute
University of Potsdam
Potsdam, Germany
lukas.ehrig@student.hpi.de

Daniel Atzberger
Hasso Plattner Institute
University of Potsdam
Potsdam, Germany
daniel.atzberger@hpi.de

Benjamin Hagedorn
Hasso Plattner Institute
University of Potsdam
Potsdam, Germany
benjamin.hagedorn@hpi.de

Jan Klimke
Hasso Plattner Institute
University of Potsdam
Potsdam, Germany
jan.klimke@hpi.de

Jürgen Döllner
Hasso Plattner Institute
University of Potsdam
Potsdam, Germany
doellner@hpi.de

***Abstract*—In many predictive maintenance scenarios, the costs for not accurately detecting or anticipating faults can be considerably higher than the cumulative costs for inspections or premature maintenance. However, conventional symmetric loss functions widely used in machine learning cannot reflect such different costs. In this paper, we propose a method to construct asymmetric loss functions for regression tasks that are capable to better reflect this cost imbalance during the training of machine learning models and that allow for modeling the loss function to a) precisely match the cost relation for both kinds of errors where they can be estimated, b) control the impact of outliers, and c) manage the risk of over- or underestimation of the target variable, even when exact costs for (at least) one side are not known. We demonstrate on a realistic data set that the customized asymmetric loss functions can significantly reduce the impact of overestimations of the remaining useful life and can help to take more informed decisions on maintenance planning, leading to more cost-efficient production processes.**

***Keywords*—predictive maintenance, loss functions, cost functions, machine learning, cost-sensitive, regression**

I. Introduction

Machine learning (ML) techniques are increasingly employed for fault detection or anticipation in the context of condition monitoring and predictive maintenance for technical assets (e.g., machines and plants) [1]. For this, asset and process data (e.g., power draw, vibration, noise, power consumption) serve as the basis for training and testing ML models that can be used for on-line assessment of asset condition and remaining useful life (RUL).

In many scenarios, the costs for not accurately detecting or anticipating faults can be considerably higher than the cumulative costs for inspections or premature maintenance. These higher costs result, e.g., from reduced machine efficiency (output, availability, and product quality), unplanned production downtimes, higher expenditures for corrective maintenance due to overtime payments, express charges for urgent supply of tools and material, subsequent damages of machines or their environment, or even physical harm of employees. By planning and executing maintenance activities earlier based on the asset condition, many of these costs can be avoided [2].

A. Example Use Case

Asymmetric costs can be observed not only in manufacturing. In aircraft maintenance, it is much cheaper to exchange wear parts of the turbofan early than to overhaul it completely later. Even if casualties can be averted in case of a failing turbofan, there are high costs resulting from the incurred delay of operation due to a rejected take-off or in-flight turnback.

Consider, for example, the flight compensation regulation of the EU that entitles passengers to financial compensation [3]. If an incident occurred off-station, e.g., at a remote holiday destination, the airline needs to fly in a maintenance crew and replacement material. The cost for aircraft on ground (AOG) can amount up to €925,000 per day for an Airbus A380 [4]. In case the aircraft can be returned to service, it has to wait for a new takeoff slot; the knock-on delay affects all subsequent flights (up to ten for low-cost carriers), if it cannot be replaced by another aircraft (tail swap). For lengthy maintenance, a wet lease (aircraft + crew) is required, resulting in high cost. Another factor to consider is the passenger goodwill loss, i.e. the loss resulting from reduced customer satisfaction.

These high costs explain the large amount of effort airlines undertake to maintain their fleet, with direct maintenance costs (DMC) accounting for 10 to 20 percent of the operating costs of an aircraft [5]. Due to increasing competition in the aviation market, operators are under constant pressure to increase the block hour utilization and minimize expenses. Efficiency drivers such as the reduction of backup aircraft manifest the role of predictive maintenance to prevent unexpected breakdowns. Even though the maintenance-related regulatory requirements are designed to be safe and aircraft are built with redundancy so that the breakdown of a part usually does not pose a significant risk to passengers, an unexpected failure still entails that the aircraft cannot take off again. Therefore, it may be economical for airlines to exchange parts based on their condition before it is required by regulation.

Due to the cost imbalance for early maintenance vs. engine failures as outlined above, overestimating the RUL and risking potential engine failures is much worse than planning the maintenance too early. In other scenarios where failures may not cause life-threatening situations and early predictions instead involve substantial economic burden, these costs may change and optimistic predictions would be preferable. Hence, the loss function should reflect such aspects to match specific requirements. However, conventional symmetric loss functions, such as the mean squared error (MSE) or Huber loss [6] cannot model asymmetric costs.

This work has been supported by the German Federal Ministry of Education and Research (BMBF) through grant 01IS19006 (KI-Labor ITSE).

B. Contributions

In this paper, we propose a method to construct asymmetric loss functions for regression tasks that are capable to better reflect a cost imbalance during the training of ML models. The proposed loss functions are continuously differentiable and convex, which is desirable for optimization methods based on gradient descent.

They can be parameterized to penalize overestimates more than underestimates or vice versa while being much more flexible than LINEX loss [7], which only supports stretching or shrinking the function as a whole. In contrast, with the proposed asymmetric loss functions the growth characteristic (i.e., linear, quadratic or exponential) and slope can be freely adjusted for both losses of under and overestimates separately. This allows us to model a loss function to a) precisely match the cost relation for both kinds of errors where they can be estimated, b) control the impact of outliers, and c) manage the risk of over or underestimation of the target variable (RUL), even when exact costs for (at least) one side are not known. So it can better represent real-world costs for too early vs. too late maintenance and their relation already during the training phase of a ML model.

II. Related Work

Asymmetric loss functions have been studied by economists in the context of econometric modeling and forecasting. Varian [7] introduced the LINEX loss for the prediction of real estate prices. It grows exponentially on one side and approximately linearly on the other side and has been applied to other areas, such as Bayesian estimation problems in statistics [8]. The LINEX loss can be extended to a Double-LINEX loss, which is exponential on both sides [9]. It has been shown that properties of optimal forecasts under MSE loss such as unbiasedness fail to hold in the optimal forecast under asymmetric loss [10]–[12]. A drawback of LINEX loss functions is that it is not possible to control the slope of the loss function for both sides independently.

In the context of ML, the most frequently used loss functions for regression are the (linear) symmetric mean absolute error (MAE) loss and the (quadratic) MSE loss. Both loss functions have been proposed and studied in their generalized, asymmetric variant. The use of an asymmetric linear loss is known as quantile regression; the use of an asymmetric quadratic loss is also known as expectile regression [13]–[15].

Berk [16] applies quantile regression to forecasts in a criminal justice setting to find forecasting procedures that are more sensitive to the real consequences of forecasting errors than statistical procedures relying on symmetric loss functions typically used for forecasting.

Huber [6] introduces a loss function that is quadratic (and therefore strictly convex) when the error is small but behaves like the MAE loss when the error is larger than a specified threshold. This leads to a continuously differentiable function which remedies properties of the MAE loss which are adverse for convergence while retaining the advantage of the MAE of being less sensitive to outliers than the MSE loss. Recently, Gupta et al. [17] proposed using an asymmetric version of this loss function for predicting sediment transport in hydrology.

An approach to cost-sensitive regression that does not rely on asymmetric loss functions is post hoc tuning, in which a model that has been trained without incorporating asymmetric costs is retroactively adjusted to account for those costs [18].

III. Proposed Method

All model-based supervised ML approaches rely on loss functions which are used during the model building process to evaluate candidate solutions and for optimization. The loss (or, strictly speaking, cost) function has to reduce all aspects of the model down into a single scalar value in such a way that improvements in that number indicate a better model. It must therefore capture the properties of the problem and specifies design goals for the search of an optimal solution.

The objective of regression analysis is to estimate the relationships between a dependent *target variable* and independent variables called *features*. In the context of predictive maintenance, the target variable is the RUL which specifies the amount of time a machine can be operated until it becomes undesirable to do so and maintenance is necessary, e.g., because of excessive wear and damage to the machine or quality deterioration. The residual x of a prediction is the difference of the true value of the target variable y and the model's estimation $\hat{y}$: $x = y - \hat{y}$. If the residual is positive, the RUL has been underestimated which is preferable in our scenario to an overestimated RUL (negative residual).

A. Asymmetric Loss Function

We define our asymmetric loss functions piece-wise, so that the loss for residuals that are greater than 0 differs from those smaller than 0:

$$\mathcal{L}(x) = \begin{cases} \ell_l(x), x < 0 \\ \ell_r(x), x \geq 0. \end{cases} \tag{1}$$

As ℓ_l and ℓ_r can be examined independently, we will focus on ℓ_r; the left-hand side can be treated analogously.

To ensure convexity and continuous differentiability, we construct the loss function in a way, that $\ell_r(x)$ is quadratic for small residuals, i.e. $\ell_r(x) = \alpha_r x^2$ up to a point denoted as θ_r, and append another function f featuring the desired growth characteristic at $\theta_r > 0$, unless the desired growth is quadratic. This function f needs to be attached in a way that (1) $f(\theta_r) = \alpha_r \theta_r^2$ and (2) $f'(\theta_r) = 2\alpha_r\theta_r$.

In this work, we consider two types of functions for the attachments: linear and exponential functions. In the linear case, this holds for $f(x) = \alpha_r\theta_r(2x - \theta_r)$.

In the case of an exponential loss, the corresponding attachment is $f(x) = \alpha_r\theta_r(\theta_r + 2\psi_r(e^{\frac{x-\theta_r}{\psi_r}} - 1))$, where $\psi_r > 0$ is a parameter allowing to control the growth of the exponential function. So, if the right-hand side ℓ_r of the loss function $\mathcal{L}$ shall be exponential (Fig. 1):

$$\ell_r(x; \theta_r, \alpha_r, \psi_r) = \begin{cases} \alpha_r\theta_r(\theta_r + 2\psi_r(e^{\frac{x-\theta_r}{\psi_r}} - 1)), x \geq \theta_r \\ \alpha_r x^2 \qquad\qquad\qquad\qquad\quad , x < \theta_r. \end{cases} \tag{2}$$

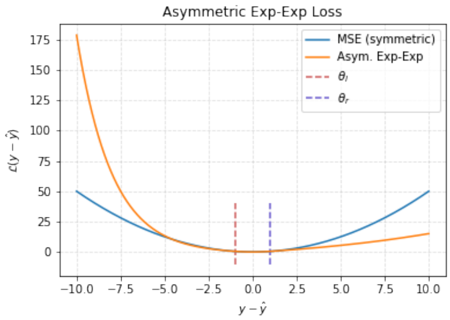


Fig. 1. An asymmetric loss function with exponential growth on both sides in contrast to the symmetric MSE loss.

Considering only three options (linear, quadratic and exponential growth) for ℓ_l and ℓ_r respectively, nine families of loss functions (including the asymmetric Huber loss) can be constructed that are parameterized by up to six parameters.

B. *Weighting*

In predictive maintenance, it is often much more important to be precise about the RUL near the end of useful life. An overestimation of RUL by, e.g., 10 flight cycles in the case of turbofan engines is very serious when only a few cycles remain and immediate action is crucial. In contrast, if dozens of flight cycles remain, the impact of inaccurate estimates on decision making is negligible. However, this is not reflected in how the loss functions usually are computed. In both cases, the residual is -10. Hence, both instances have the same loss function value and are given equal importance during the optimization of the model.

As it is even more difficult to predict further into the future than predicting at a time closer to the end-of-life, it can be expected that the residuals are larger for the less important instances than important ones, leading to a misdirected focus during the model optimization.

Hence, instance weighting should be taken into consideration when designing a predictive maintenance solution. A model-independent way of including instance weights in the training procedure is to factor it into the calculation of the loss function by weighting the residual: $x = \omega(y, \hat{y})$. A plausible choice for ω is to return the relative residual rather than the absolute residual $(y - \hat{y})$:

$$\omega(y, \hat{y}) = \frac{y - \hat{y}}{\tilde{y}}, \tag{3}$$

where $\tilde{y}$ is y if $y \neq 0$ and some constant $0 < c < 1$ otherwise. In reality, the real importance of a training instance might have to be determined by a more sophisticated weighting function. E.g., in the case of turbofan engines, the same relative overestimation of the true RUL is worse for a small RUL than when it is very large.

IV. Experiments & Discussion

To evaluate the impact of choosing an asymmetric loss instead of standard symmetric loss functions, we trained gradient boosted decision trees (GBDTs) [19] on the NASA turbofan engine degradation data set [20], [21] with symmetric and asymmetric losses. The data consists of a run-to-failure simulation with multiple noisy multivariate time series from different engines of unknown initial degrees of wear and manufacturing variation. Each time step corresponds to one cycle (flight) and comprises snapshot measurements of 21 different sensors. The task is to predict the number of remaining cycles of the turbofan engine until it fails.

To train the GBDTs we construct the training data by splitting each of the multivariate time series into windows of different lengths and starting points and label them with the remaining number of cycles, i.e., the number of following time steps. For each window, we extract features using the FRESH algorithm [22] and select the 500 most relevant features with univariate linear statistical tests. Each model has been trained with 400 boosting iterations, a different loss function and is evaluated with 5-fold cross validation. The influence of the asymmetry of the loss functions on the error distribution is shown in Fig. 2 for three different loss functions with exponential loss and relative weighting on both sides; corresponding summary statistics can be found in Tab. I.

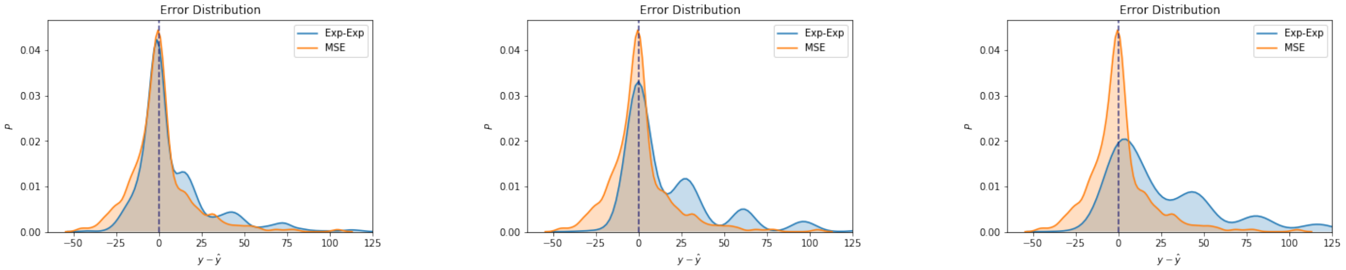
All three models trained with asymmetric loss functions do not only overestimate the RUL less often but also less severe than models trained with MSE loss. The asymmetric loss functions also overestimate the RUL, but mostly by small amounts, which are not so severely penalized by the loss function. The opposite applies to underestimates which are less penalized by the loss functions and more acceptable in our scenario.

TABLE I. Summary statistics and loss function values (rounded average of 5 CV-rounds) for different loss functions.

	MSE	Exp_1	Exp_2	Exp_3
MAE	11.98	14.15	18.98	26.77
$\sqrt{\text{MSE}}$	18.32	23.87	32.48	42.35
% Over- / Underest.	55 / 45	47 / 53	30 / 70	22 / 78
Max —"—	40.4 / 84.5	28.5 / 106.1	16.8 / 127.9	14.9 / 146.7
Avg —"—	10.0 / 14.5	6.3 / 21.2	3.4 / 25.9	1.5 / 33.7
Median —"—	7.4 / 7.8	3.4 / 14.4	2.3 / 20.8	0.44 / 22.1

The MAE is larger for the asymmetric loss functions than for the MSE loss. A reason for this is weighting, which leads the model to not giving attention to training instances with large MAE because of a large RUL. However, if only instances with short RULs of 30 cycles or less are considered, the MAE of the models trained with the asymmetric loss functions is smaller. The MAE accounts neither for asymmetric costs nor for different instance importance. Experimental results show that the impact of asymmetric loss functions is significant, even for relatively small asymmetries. In experiment Exp_1, the average and median overestimation could approximately be halved, with only a marginal increase of the MAE over all instances. The MAE for the most relevant instances with a RUL of 30 cycles or less even decreased from 3.96 (MSE loss) to 2.71 (Exp_1 loss). Depending on the risk aversion, the parameters can be adjusted in a way that severe overestimations become very unlikely, with the median overestimation of models trained with loss Exp_3 of 0.44 being insignificant.

Experiments showed that the asymmetry and parameterization of the loss function affects the appropriate param-



(a) Exp_1 (b) Exp_2 (c) Exp_3

Fig. 2. The error distribution for estimations from models trained with asymmetric loss functions with increasing asymmetry from (a) to (c). For comparison, the error distribution of models trained with the standard MSE loss is also shown.

eterization of the optimizer. Therefore, the learning rate should be adjusted so that Gradient Descent takes step sizes corresponding to the slope of the loss function.

V. Conclusion

In this paper, we introduced a method to construct loss functions that can be adapted to problems with asymmetric costs and different instance importance needed in fields such as predictive maintenance. We motivate asymmetric costs with a case study from commercial aviation and demonstrate that the number and impact of overestimates of the RUL can be significantly reduced and controlled according to the respective risk aversion. This leads to ML models that reflect the real business situation far better than models trained with standard symmetric loss functions.

Future work includes studying the convergence behavior of optimization methods for different degrees of asymmetry, methods for choosing the right parameters such as the learning rate to ensure a fast convergence, examination the impact of custom asymmetric loss functions on different ML models, and to find and evaluate different weighting schemes.

References

[1] C.-W. Chang, H.-W. Lee, and C.-H. Liu, "A review of artificial intelligence algorithms used for smart machine tools," *Inventions*, vol. 3, no. 3, p. 41, 2018.

[2] A. Rastegari and M. Bengtsson, "Cost effectiveness of condition based maintenance in manufacturing," in *2015 Ann. Reliability and Maintainability Symp. (RAMS)*, 2015, pp. 1–6.

[3] "Regulation (ec) no 261/2004 of the eur. parl. and of the council," *Official Journal of the European Union*, vol. L 46/1, 2004.

[4] Keeping the aircraft flying – aircraft on ground (aog) services. Accessed: 2020-03-31. [Online]. Available: https://www.dhl.com/en/logistics/industry_sector_solutions/aerospace_logistics/aircraft_on_ground.html

[5] U. PeriyarSelvam, T. Tamilselvan, S. Thilakan, and M. Shanmugaraja, "Analysis on costs for aircraft maintenance," *Advances in Aerospace Science and Applications*, vol. 3, no. 3, pp. 177–182, 2013.

[6] P. J. Huber, "Robust estimation of a location parameter," *J Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, 1964.

[7] H. R. Varian, "A bayesian approach to real estate assessment," *Studies in Bayesian Econometrics and Statistics*, pp. 195–208, 1975.

[8] A. Zellner, "Bayesian estimation and prediction using asymmetric loss functions," *J Amer. Stat. Assoc.*, vol. 81, no. 394, pp. 446–451, 1986.

[9] C. W. Granger, "Outline of forecast theory using generalized cost functions," *Spanish Economic Review*, vol. 1, no. 2, pp. 161–173, 1999.

[10] C. Gravel, "Prediction with a generalized cost of error function," *J Operational Research Society*, vol. 20, no. 2, pp. 199–207, 1969.

[11] P. F. Christoffersen and F. X. Diebold, "Optimal prediction under asymmetric loss," *Econometric Theory*, vol. 13, no. 6, p. 808–817, 1997.

[12] A. J. Patton and A. Timmermann, "Properties of optimal forecasts under asymmetric loss and nonlinearity," *J Econometrics*, vol. 140, no. 2, pp. 884–918, 2007.

[13] R. Koenker and G. Bassett, "Regression quantiles," *J Econometrica*, vol. 46, no. 1, pp. 33–50, 1978.

[14] W. K. Newey and J. L. Powell, "Asymmetric least squares estimation and testing," *J Econometrica*, vol. 55, no. 4, pp. 819–847, 1987.

[15] L. Liao, C. Park, and H. Choi, "Penalized expectile regression: an alternative to penalized quantile regression," *Annals of the Institute of Statistical Mathematics*, vol. 71, no. 2, pp. 409–438, 2019.

[16] R. Berk, "Asymmetric loss functions for forecasting in criminal justice settings," *J Quant Criminol*, vol. 27, no. 1, pp. 107–123, 2011.

[17] D. Gupta, B. B. Hazarika, and M. Berlin, "Robust regularized extreme learning machine with asymmetric huber loss function," *Neural Computing and Applications*, pp. 1–28, 2020.

[18] H. Zhao, A. P. Sinha, and G. Bansal, "An extended tuning method for cost-sensitive regression and forecasting," *Decision Support Systems*, vol. 51, no. 3, pp. 372 – 383, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167923611000285

[19] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *J Annals of Statistics*, pp. 1189–1232, 2001.

[20] A. Saxena and K. Goebel, "Turbofan engine degradation simulation data set," *NASA Ames Prognostics Data Repository*, 2008. [Online]. Available: https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository

[21] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage propagation modeling for aircraft engine run-to-failure simulation," in *2008 Int. Conf. on Prognostics and Health Management*, 2008, pp. 1–9.

[22] M. Christ, A. W. Kempa-Liehr, and M. Feindt, "Distributed and parallel time series feature extraction for industrial big data applications," *arXiv preprint arXiv:1610.07717*, 2016.